**第二次汇报**

程序见：my\_project\_text\_classification\_4

1. 词向量和模型

max\_features=4000, lowercase=False, ngram\_range=(1,2)

根据词频做词向量(countVector)、TF-IDF做词向量(10折交叉验证)

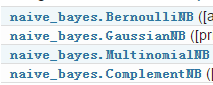
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方式 | KNN | 贝叶斯 | SVM | 随机森林 | 神经网络 |
| 词频(CV) | 0.57 | 0.843 | 0.814 | 0.5355 | 0.835 |
| TF-IDF | 0.819 | 0.768 | 0.8524 | 0.767 | 0.844 |

**调参:**

max\_features=20000, lowercase=False, ngram\_range=(1,3)，11折交叉验证

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方式 | KNN | 贝叶斯 | SVM | 随机森林 | 神经网络 |
| 词频(CV) | 0.566 | 0.8564 | 0.8120 | 0.8080 | 0.8470 |
| TF-IDF | 0.8214 | 0.7628 | 0.8522 | 0.8091 | 0.8602 |

**文本的卷积神经网络的概念大致理解了，现在还在看相关的代码。**

min\_samples\_split=2, min\_samples\_leaf=1

****